Feminist Curation of Text for Data-centric AI

Marion Bartl and Susan Leavy
Insight SFI Research Centre for Data Analytics
School of Information and Communication Studies
University College Dublin
marion.bartl@insight-centre.org, susan.leavy@ucd.ie

Abstract

Language models are becoming increasingly central to artificial intelligence through their use in online search, recommendation engines and language generation technologies. However, concepts of gender can be deeply embedded in textual datasets that are used to train language models, which can have a profound influence on societal conceptions of gender. There is therefore an urgent need for scalable methods to enable the evaluation of how gender is represented in large-scale text datasets and language models. We propose a framework founded in feminist theory and feminist linguistics for the assessment of gender ideology embedded in textual datasets and language models, and propose strategies to mitigate bias.

1 Introduction

Language models that underlie many artificial intelligence (AI) systems are commonly trained on text from sources such as Google Books [11] or the Common Crawl[1]. These datasets incorporate a variety of social biases, among them gender bias, and when used as training data, can result in discriminatory AI systems [18, 8]. Gender bias is expressed in language in various forms, for instance in the use of stereotypical or sexist language or imbalances in the distribution of mentions of people of different genders. Gender bias in text can be amplified by language-based AI systems and thus cause representational and allocational harms to women and non-binary people [2]. As a response, calls for better data evaluation and curation, as opposed to fixing trained systems, have increased. However, due to the sheer size of datasets that are used to train language models like BERT [5] or GPT-2 [16], as well as a lack of reliable and efficient methods of measuring gender bias in the training data directly, data evaluation remains a challenge. In this paper, we therefore present practical strategies for the analysis of gender in large-scale English text corpora informed by feminist linguistics. Our methods are scalable to large corpora and provide opportunities for feminist data curation.

2 Feminist linguistics and gender bias in NLP

While research on gender bias is relatively nascent within the Natural Language Processing (NLP) community [3, 20], in the field of feminist linguistics, gender bias in language has been continuously researched for almost 50 years [9]. As a result, gender-inclusive language strategies have been developed in order to counter male-centric and sexist language [6, 14]. However, as Leavy et al. [10] and Rogers [17] observe, findings from feminist linguistics have not yet informed gender bias mitigation efforts in mainstream NLP. Especially in light of recent calls for data-centric bias mitigation [17, 1], feminist linguistics can provide deeper insights into the representation of gender in these data and curate high-quality, gender-inclusive datasets. Previous data-centric approaches, such as Counterfactual Data Augmentation (CDA) [12, 13], have already proven useful for mitigating gender

[1] https://commoncrawl.org/

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

stereotypes. Still, CDA is dependent on manually curated lists of words that largely only support binary gender, which can prove exclusionary to trans and non-binary people [7].

3 Practical strategies for automatic data evaluation and curation

We first propose two working examples of assessing representations of gender in English data, which are indicative of how biases in trained models are informed. Subsequently, we outline how these strategies can be used for data-centric mitigation of gender bias.

Lexical gender of words Exploring the distribution of words that have lexical gender (such as *girlfriend* or *policeman*) can provide insights into whether masculine or feminine gender expressions predominate in a corpus. However, in English, lexical gender information is not always deducible from morphological features, such as the suffix *-man* in *policeman*. Rather, it is 'hidden', such as in the word *wife*, whose feminine lexical gender is not expressed morphologically. In order to deduce lexical gender for any given word without resorting to a lookup table, such as used in CDA [12], we propose to query the dictionary definitions of words for gender information. When implementing this strategy, a possible challenge would be to match the sense of the target word with the correct sense in the dictionary definition. Moreover, finding gender-neutral replacements for gendered words, such as *police officer* instead of *policeman* [6], can help to assess gender-inclusive language strategies in a text and thus whether its authors support a binary conceptualization of gender. However, since gender-neutral alternatives are not necessarily included in dictionary definitions of words with lexical gender, finding gender-neutral variants remains a look-up problem for now.

Epicene coreference While the previous strategy targets individual words, our second strategy aims at syntactic constructions, specifically pronoun coreference. If the gender of an antecedent is unknown, the gender of the referencing pronoun (epicene pronoun) can provide insights into whether gender stereotypes are at play (e.g. referencing a nurse of unknown gender by the pronoun *she*), whether a corpus follows a male-as-norm viewpoint (e.g. using epicene *he*), or whether gender-inclusive language is used (e.g. epicene *they*). We propose to use a coreference resolution algorithm to detect coreference clusters, analyze the alignment of antecedent and pronoun gender, and thus deduce whether gender-inclusive or gender-biased language strategies are used. Challenges for this approach are the dependence on accurate coreference resolution as well as antecedent gender detection, for which the previous strategy of dictionary-based detection could be applied.

Data-centric gender bias mitigation Once an overview on the representation of gender has been obtained, we can then specifically target male-as-norm language (e.g. *policeman*, epicene *he*), replace it with more inclusive forms (e.g. *police officer*, epicene *they*), and, after training AI models on these more gender-inclusive texts, measure the impact of gender-based data curation on gender bias. Since there exists a positive cognitive effect of using inclusive vs. sexist language [15, 19, 4], this effect could also be absorbed by AI systems trained on gender-inclusive text. Another auxiliary advantage of gender-inclusive training data is that systems learn to incorporate gender-inclusive phrasing, i.e. language that does not assume gender or the gender binary from the start, which can be especially useful in text generation systems. On the other hand, neutralization strategies for gendered words and coreference clusters might not be enough to prevent e.g. contextualized word representations to carry latent gender information, since inferences within language models like BERT [5] or GPT-2 [16] are drawn from word co-occurrences within billions of tokens of text.

4 Conclusion

Feminist linguistics can provide the necessary theoretical and practical background to facilitate the shift from model-centric to data-centric bias mitigation in NLP. On one hand, gendered language detection can be useful to assess data quality in terms of prevalence for a certain gender and gender-inclusiveness. On the other hand, automatically curating a corpus based on principles of gender-inclusive language could mitigate gender stereotypes for trained models on these data and provide better representation of people with non-binary gender identities. Thus, feminist linguistic interventions present interesting avenues for future research in text-based AI systems. We believe that it is time to integrate 50 years of research on sexism in language into the computational models of today.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 12/RC/2289_P2. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

[1] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[2] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, 2020.

[3] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357, 2016.

[4] Soledad de Lemus and Lucía Estevan-Reina. Influence of sexist language on motivation and feelings of ostracism (la influencia del lenguaje sexista en la motivación y el sentimiento de ostracismo). *International Journal of Social Psychology*, 36(1):61–97, 2021. doi: 10.1080/02134748.2020.1840230.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Francine Harriet Wattman Frank, Paula A. Treichler, and Modern Language Association of America. Commission on the Status of Women in the Profession. *Language, gender, and professional writing: theoretical approaches and guidelines for nonsexist usage*. Commission on the Status of Women in the Profession, Modern Language Association of America, New York, 1989. ISBN 9780873521789;087352179X;0873521781;9780873521796;.

[7] Abbie E. Goldberg and Katherine A. Kuvalanka. Navigating identity development and community belonging when "there are only two boxes to check": An exploratory study of nonbinary trans college students. *Journal of LGBT Youth*, 15(2):106–131, 2018. doi: 10.1080/19361653.2018.1429979. URL `https://doi.org/10.1080/19361653.2018.1429979`.

[8] Wei Guo and Aylin Caliskan. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133, 2021.

[9] Robin Lakoff. Language and woman's place. *Language in society*, 2(1):45–79, 1973.

[10] Susan Leavy, Eugenia Siapera, and Barry O'Sullivan. Ethical data curation for ai: An approach based on feminist epistemology and critical theories of race. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 695–703, 2021.

[11] Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pages 169–174, 2012.

[12] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer, 2020.

[13] Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275, 2019.

[14] Sara Mills. *Feminist stylistics*. Routledge, 2002.

[15] Janice Moulton, George M Robinson, and Cherin Elias. Sex bias in language use:" neutral" pronouns that aren't. *American psychologist*, 33(11):1032, 1978.

[16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[17] Anna Rogers. Changing the world by changing the data. *CoRR*, abs/2105.13947, 2021. URL `https://arxiv.org/abs/2105.13947`.

[18] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, 2019.

[19] Jane G Stout and Nilanjana Dasgupta. When he doesn't mean you: Gender-exclusive language as ostracism. *Personality and Social Psychology Bulletin*, 37(6):757–769, 2011.

[20] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, 2019.